2021 Flash Flood and Intense Rainfall (FFaIR) Experiment: Executive Summary

Sarah Trojniak, James Correia Jr., and Jim Nelson *November 2021*

Introduction

The Flash Flood and Intense Rainfall (FFaIR) experiment was held for the ninth time this summer and for the second time it was completely virtual. The experiment took place from June 21 to July 23, 2021. The FFaIR experiment is part of the Hydrometeorology Testbed (HMT) at the Weather Prediction Center (WPC) and focuses on evaluating the utility of experiment guidance and products in the forecasting process for heavy rainfall and flash flooding. The experiment itself is traditionally held in person in the WPC HMT Collaboration Room, but like last year, due to the continuing pandemic, the experiment was once again completely virtual. Even in a virtual setting, FFaIR continues to succeed at bringing together meteorologists across the weather enterprise, ranging from National Weather Service (NWS) forecasters to academia to model developers, to work together in a pseudo-operational setting towards increasing the skill of heavy rainfall and flash flood forecasts. Evaluation of experimental guidance in the FFaIR experiment helps navigate the NWS's process of transitioning new tools and products into operations (referred to as the R2O process) by identifying what may or may not be ready to be used operationally across the NWS.

Data

The guidance that was evaluated in FFaIR can be found in Table 1. This year the majority of the experimental guidance was centered around various convecting allowing models (CAMs) and ensemble configurations at use the FV3 model core, which will eventually be the biases for

all NWS models and ensembles; the whole convective allow suite, once implemented, will be referred to as the Rapid Refresh Forecast System or RRFS. This included three different deterministic configurations provided by the Environmental Modeling Center (EMC), which use LAM (limited area model) as an identifier for their model. One configuration, referred to as the RRFS1, was provided by the Global Systems Laboratory (GSL). Additionally, EMC and GSL, along with collaboration with the National Severe Storms Laboratory (NSSL), provided a RRFS Ensemble run in the cloud (referred to as RRFSCE). Finally, the Center for Analysis and Prediction of Storms (CAPS) provided their Storm Scale Ensemble Forecast (SSEF) system and four of the members from the SSEF were evaluated as deterministic models.

In addition to the various configurations of FV3-CAMs from GSL, EMC and CAPS, the Colorado State University (CSU) Machine-Learning Prediction (MLP) team once again provided numerous versions of their MLP "first-guess" Excessive Rainfall Outlook (ERO) products, all of which are CAM based. One of their MLP EROs has already been transitioned to operations (GEFS-based) and they are now working towards unlocking the details provided by CAMs to be used as guidance for WPC forecasters when creating EROs. The MLP ERO suite included four versions trained on the NSSL model (NSSL2, NSSL3, NSSL4 and NSSL5), one trained on the HRRR model and one that is a blend (referred to as the BLEND ERO) of the NSSL2, HRRR, and GEFS EROs.

Goals

All of the experimental goals and objectives are listed below:

- Evaluate the usefulness of operational and experimental products from high resolution convective-allowing deterministic and ensemble models for forecasting extreme rainfall and flash flood events, with the main focus on the Day 1.
- Collect more information on the prolific forecasting of grid point storms in the FV3-CAMS that were identified in the 2020 FFaIR Experiment (see the 2020 FFaIR Final Report). Identify aspects of the cells such as size, rain rate, timing, weather patterns, etc.
- Focus on the guidances' ability to correctly forecast precipitation events exceeding various thresholds such as 2, 4 and 6 inches.
- Evaluate models' and ensembles' timing of precipitation onset, progression, and end during a 6 h time period.
- Evaluate CSU MLP for the Day 1 ERO, which this year focuses on using CAM models for the forecasts.
- Explore using Average Recurrence Intervals (ARI) as the base for an excessive rainfall outlook. These will be referred to as an ARI-ERO.

Table 1: *Summary of the experimental guidance evaluated in the 2021 FFaIR Experiment.*

Summary of Models, Ensembles and Products for 2021 FFaIR	
EMC	Three versions of the FV3-Limited Area Model (LAM): FV3-LAM FV3-LAMX FV3-LAMDAX
EMC/GSL	Rapid Refresh Forecast System (RRFS) Ensemble run on the Cloud: RRFS-Cloud or RRFSCE
GSL	RRFS1
OU-CAPS	Storm Scale Ensemble Forecast (SSEF)
OU-CAPS	Four Members of the SSEF as deterministic models: SSEF Control Member RRFS-like Member HRRR-like Member WoFS-like Member
CSU	First Guess ERO Fields from: HRRR BLEND NSSL-sptavg -> NSSL2 NSSL-sptavg-landsea-mask -> NSSL3 NSSL-sptavg-landsea-mask-params -> NSSL4 NSSL-tempavg -> NSSL5

Activites

Three forecasting activities were done in FFaIR this year to help evaluate the utility of the guidance provided. Two of the activities, the creation of an ERO and the Maximum Rainfall and Timing Product (MRTP), were part of the activities in the 2020 FFaIR Experiment. New this year was the creation of the ARI-ERO. Both the ERO and ARI-ERO are Day 1 products, valid 16 UTC to 12 UTC just like the operational ERO. The ERO identifies the probability of rainfall exceeding FFG within 40 kilometers (25 miles) of a point will occur. The risk categories are: Marginal (5-10%), Slight (10-20%), Moderate (20-50%), and High (>50%). The ARI-ERO was designed by the FFaIR team to try and assess the utility of ARI exceedances in identifying areas where heavy rainfall would occur and might lead to flash flooding. The product identities were in any given six hour time interval within the valid time of the product (16 UTC to 12 UTC) rainfall has a 75% chance of exceeding the 1, 2, 5, and 10 year ARI. Therefore the ARI categories are: 1yr AIR 6h, 2yr, ARI 6h, 5yr ARI 6h, and 10yr ARI 6h. An example of an ERO and ARI-ERO created by the participants, along with verification information, can be seen in Fig. 1.

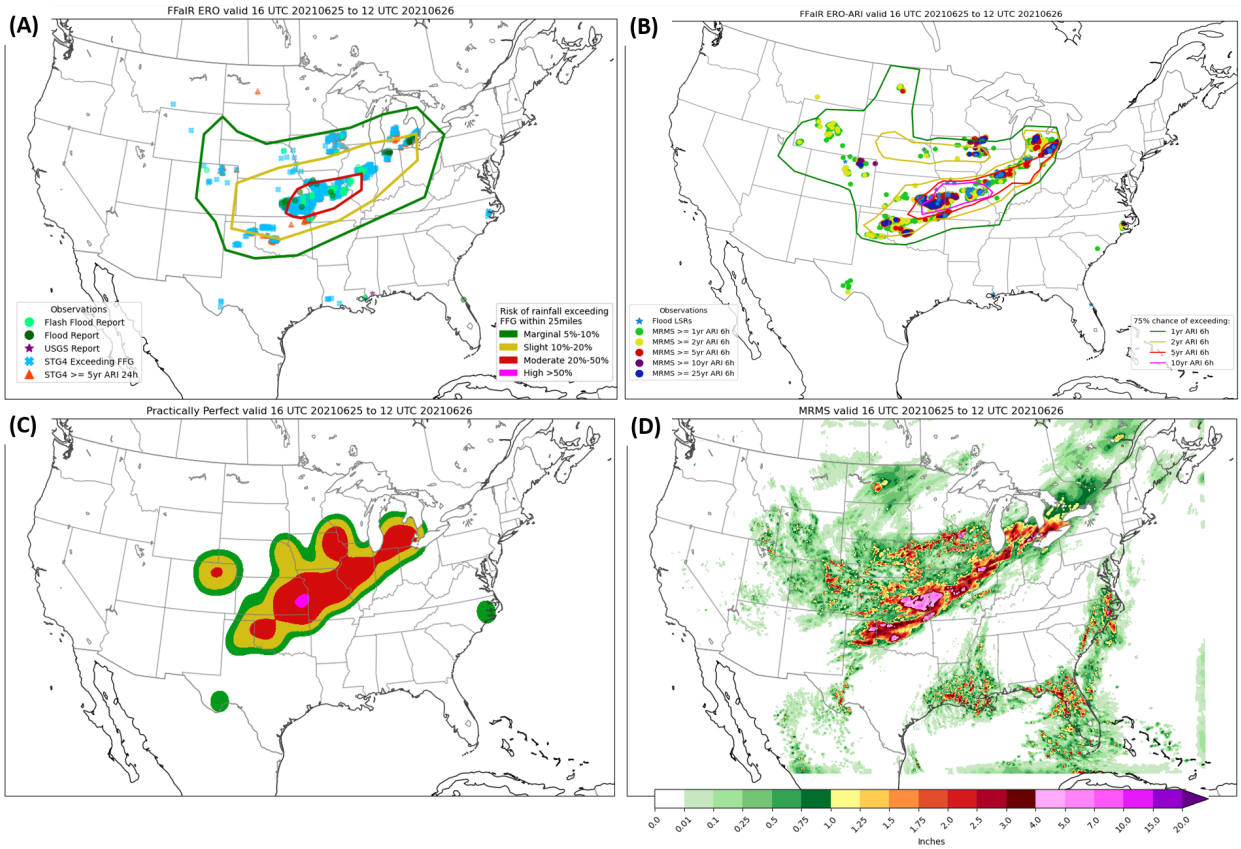

Figure 1: *(A) FFaIR ERO with UFVS observations plotted (see legend), (B) FFaIR ARI-ERO, (C) ERO practically perfect verification and (D) MRMS valid 16 UTC 25 June to 12 UTC 26 June 2021.*

The MRTP is an individual forecast activity that is valid for a 6h period and over a specific region, both of which are determined each day by the participants collectively. The product itself consists of three aspects, the creation of a QPF forecast, the evaluation of an assigned model or ensemble and the completion of a survey. Participants were not required to use the assigned model or ensemble for their forecast but they were required to analyze the guidance and state in their survey why they did or did not use what they were assigned. The survey consisted of asking participants things like what they thought the maximum rainfall total will be, what ARI would be exceeded, how long the rainfall would occur, if flooding would occur and its intensity, and what models/ensembles they found useful and which they didn't. They were also asked to provide a summary of the forecast challenge and what they found useful in the guidance they were using. The full survey can be found in Appendix C of the 2021 FFaIR Final Report. For the drawing of the MRTP, the participants had the option to draw contours for six hour rainfall totals of 1 inch, 2 inches, 3 inches, and 4 inches and to identify where they thought the highest rainfall total would occur. They also could highlight where they thought rainfall rates would exceed 1in/h. An example of two MRTPs can be seen in Fig. 2.

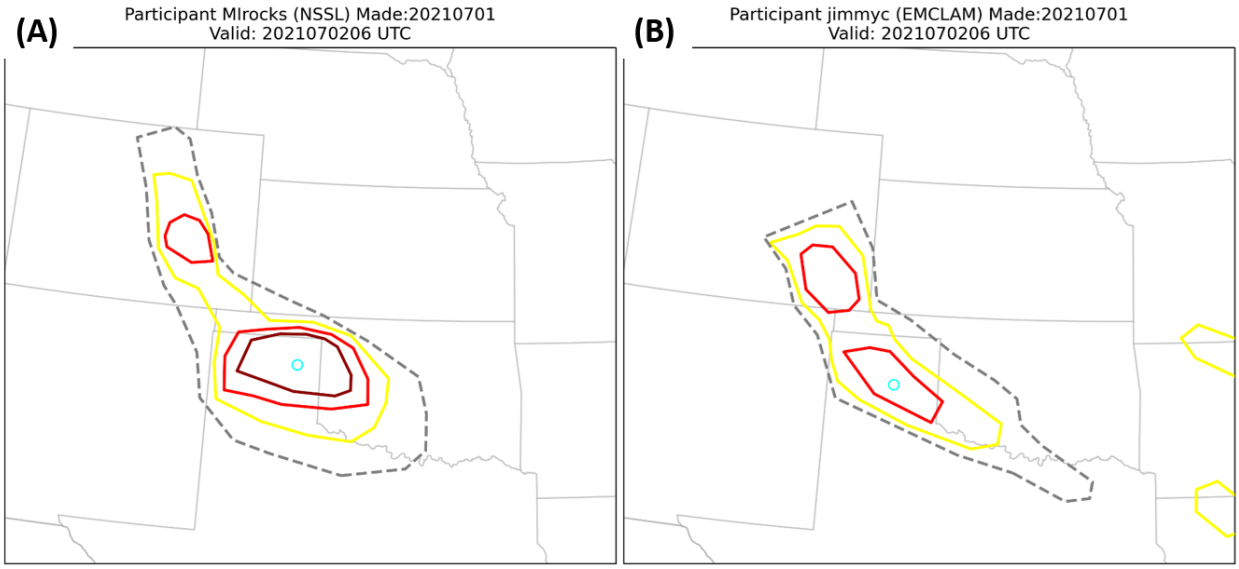

Figure 2: *Example of two MRTPs issued on July 01, valid from 00 UTC to 06 UTC July 02, 2021. Usernames for the MRTPs: (A) MIrocks and (B) jimmyc. Contours: yellow - 1 in., red - 2 in., dark red 3 in., purple - 4 in., and dashed gray - 1in/h rainfall rates. The blue circle is the forecasted location of maximum rainfall.*

Summary and Research-to-Operations Recommendations

The main findings and recommendations are summed up in the following bullet points. Table 2 identifies what the transition recommends are for the guidance.

- EMC provided two models, the **LAM** and **LAMX**, that were identical aside from the domain they were run on. The LAM was run on a domain similar to the HRRR CONUS domain while the LAMX was run on the RRFS North American domain. The goal of this was to determine if the larger domain had a significant impact on the forecast. Both subjective and objective evaluation of the LAM and LAMX show that there was little difference in the QPF forecasts between these two models. **Therefore it is recommended that EMC move forward with running their LAMs on the larger, Northern American grid.**

- EMC also provided a FV3 configuration that included data assimilation, **LAMDAX**. Subjectively this was liked less than the LAM and LAMX by the participants. However, contingency table metrics suggest its performance, at least for the 00z run, is similar to the other two LAMs, with a slightly lower wet bias. **It is recommended that the data assimilation methodology for the LAMs continue to be developed.**

- The **RRFS1** provided by GSL was the least liked FV3-CAM by the participants. Participants often noted that the QPF footprint/storm mode did not resemble observations. Its 00z cycle was the lowest performing FV3-CAM evaluated during FFaIR at both the half inch and one inch 24h QPF thresholds. The 12z cycle's

performance was more comparable to the LAM and LAMX, though the RRFS1's wet bias was greater than the LAM/LAMX's wet bias for this cycle. **Recommended for continued development.**

- At the half inch and one inch 24 h QPF threshold, the FV3-CAMs have a similar bias to the operational models, with the SSEF members, RRFS1 and LAMDAX all having a slight dry bias for the half inch threshold. At the higher end 24h QPF thresholds (2+ inches) all FV3-CAMs have a wet bias greater than the HRRR. At two inches the bias is similar to the NAMnest but at three inches the wet bias seen in the FV3-CAMs is greater than the NAMnest wet bias. In some instances the wet bias approached 5 from some models. This suggests that the FV3 models generally underforecast the occurrence of rainfall (dry bias at low thresholds) but when they do forecast rainfall they overforecast the magnitude of the precipitation. **Therefore it is recommended that developers work to identify what is driving the FV3-CAM's difficulty initiating precipitation but once initiating it over forecasting amounts.**

- The wet bias in the LAMs, RRFS1, and to a lesser extent the SSEF members is very evident when single cell (popcorn) convection is forecasted. The size and shape of the cells resemble grid cells. In many instances, especially for the RRFS1, when the popcorn convection was forecast, nearly every cell had hourly QPF exceeding 2 inches. For the RRFS1 specifically, there were times when the hourly QPF from these cells exceeded 9 inches; an example of this can be seen in Fig 3. **Until identification of what is driving the overproduction of rainfall in single cell convection in the FV3-CAMs, implementation of any configurations should not occur.**

- The coverage of instantaneous precipitation rates (p-rate) from the models evaluated during subjective verification was smaller when compared MRMS. However when focusing on the maximum p-rate, the SSEF CNTL member, LAMDAX and RRFS1 generally had a greater maximums than the HRRR or MRMS did. When evaluating the maximum p-rate from June 10 to July 31, the average maximum p-rate from the LAMs were roughly double the average maximum from the MRMS and HRRR. The average maximum p-rate from the RRFS1 was 3 to 4 times as great, with some maxima exceeding 150 in/hr. **The FV3- CAMs, especially the RRFS1, have a tendency to output p-rates that are much larger than observed and at times that are unrealistic. It is possible that these rates are helping to drive the large QPF values seen in the single cell convection from the models. The impact of these p-rates and what is driving such high magnitudes should be examined further.**

- Neither the SSEF or the RRFSCE outperformed the HREF. When focusing on probability of exceedance thresholds, the SSEF probabilities were extremely low. The RRFSCE probabilities were felt to be comparable to the HREF probabilities at times by the participants, but this was generally when the ensemble appeared to have a good handle on

the pattern. When this occurred, probabilities were generally high due to the low spread in the ensemble. The mean products (traditional mean, PMM, and LPMM) from the SSEF, for the most part, had a slightly higher CSI than the respective means of the RRFSCE. The RRFSCE means had a similar bias to the HREF while the SEFF means consistently had a lower bias than the RRFSCE and HREF. Interestingly, the PMM, which is known for having a high bias, from the SSEF at a half inch and an inch had a dry bias. Therefore, it is likely that the SSEF overall is less likely to forecast precipitation in general. **It is recommended that both the SSEF and RRFSCE should continue with development. The RRFSCE appears to suffer from underdispersion and ways to introduce greater spread into the model should be examined.**

- All of the NSSL-based CSU "first-guess" EROs, except NSSL5, performed similarly to one another. The NSSL2 was the favorite of the participants while NSSL4 had a slight edge over the other ERO models in the statistical analysis. Based on frequency of issuance, the NSSL4 is more likely to forecast higher ERO risk categories than the other models and the FFaIR ERO. However, this is not necessarily a negative as participants and WPC forecasters have commented that they prefer the "first guess" to have a slight high bias since it alerts them to where they need to focus their analysis on. **It is recommended that CSU continues development on the NSSL EROs, especially the configurations for NSSL2 and NSSL4.**

- The HRRR-based ERO had the lowest performance during FFaIR, both subjectively and objectively. However this was likely in part due to the Monsoon being the dominant forecast challenge during the last two weeks of FFaIR. As discussed in Section 4.3.1 of the Final Report, the HRRR model QPF had a dry bias across the southwest during the Monsoon. Numerous things, such as lack of observations across Mexico to ingest into the HRRR's DA, help lead to the low bias across the region. This likely impacted the performance of the HRRR-based ERO model across the southwest. Additionally, the HRRR-based ERO has a short training period that does not include an active Monsoon season. **Therefore it is recommended that the HRRR-based ERO training period is adjusted to include this year's warm season, and thus the Monsoon.**

- The CSU BLEND ERO was the most liked "first guess" ERO by the participants during the first half of FFaIR. During the second half, its performance was hampered by the poor performance of the HRRR-based ERO, which is one of the three ERO models used to create the BLEND ERO. **The BLEND ERO should continue to be refined and perhaps re-evaluation of how the weights of each ERO model are determined in the creation of the BLEND ERO.**

Table 2: *Research to Operations Transition Metrics for the 2021 FFaIR Experiment.*

Models, Ensembles and Products Evaluated	Recommended for transition to operations	Recommended for further development and testing	Rejected for further testing	Provider/Funding Source
LAMX		X		EMC
LAMDAX		X		
RRFS1		X		GSL
RRFSCE		X		EMC/GSL/NSSL
SSEF		X		OU/CAPS
CSU-ML Day 1 ERO NSSL2/NSSL3/ NSSL4/NSSL5		X		CSU/JTTI
CSU-ML Day 1 ERO HRRR		X		
CSU-ML Day 1 ERO BLEND		X		

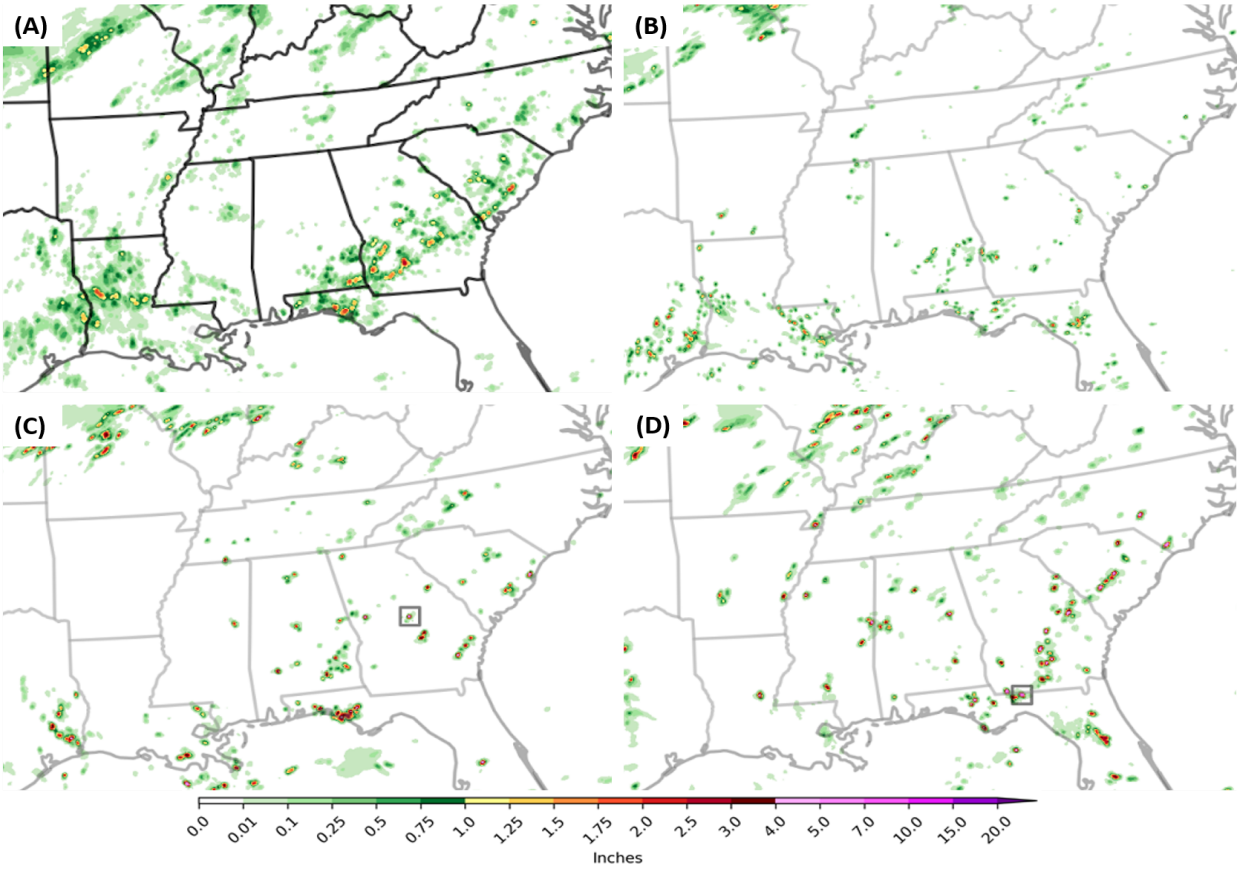

Figure 3: *1h (A) MRMS QPE and (B) HRRR, (C) LAM, and (D) RRFS1 QPF valid 21 UTC 15 July 2021. The grey box indicates the location of the model maximum across the CONUS; neither the MRMS or HRRR CONUS maximum was located in the southeast but both LAM (6.01") and RRFS1 (9.14") were.*